Including Evidence-Based Assertions in Encoded Archival Standards

First draft, submitted by Mark Custer
[see also https://docs.google.com/document/d/1y4DqqHYObO9M1qHHRjOFAQB5_9YSSl9iWiY0lcaxf7I/edit#heading=h.e5omqjh65oso]

Background

I have written this discussion paper in response to Issue #43, titled “Assertion Description (or Assertion Control)”. Issue #43 contains a feature request that was submitted by Daniel Pitti, on behalf of the Social Networks and Archival Context project, to update EAC-CPF so that the EAC-CPF schema can accommodate supplementary metadata to cite sources that provide support for why particular assertions exist in the EAC-CPF record.

In the feature request, Daniel asks that the EAC-CPF schema be updated so that the following six additional metadata fields can be encoded:

- R1: Agent who adds an assertion
- R2: Date that the agent adds (or revises) the assertion
- R3: A citation to the source that supports the assertion
- R4: A description of the evidence found in the source
- R5: Rules used in formulating the assertion
- R6: Catch-all note to accomodate any other information about the assertion

What follows is an exploration of what it might look like to include evidence-based assertions within EAC. In his paper, which will be discussed at an in-person EAC meeting on August 1, 2019, I will first present four use cases, and then proceed to demonstrate how those use cases could be included in a revised EAC-CPF schema. After the summary, I will also provide more detail about how I arrived at the possible encoding options mentioned in this paper, and I will wrap up with a few notes about additional considerations.

Use Cases

In order to test out different possibilities for how one might encode evidence-based assertions in EAC-CPF, I have decided to define four specific use cases.

Use Case 1

Consider the fact that Langston Hughes was born on February 1, 1902. That is the birthday that Langston recognized, as well as the one listed in a number of biographies, library authority

records, and SNAC's own "constellation" record, https://snaccooperative.org/view/5460855. About a year ago, however, a number of newspaper articles were published that detailed how someone recently discovered (while searching digitized newspapers) that it is extremely likely that Langston Hughes was actually born one year earlier, in 1901.

How might you encode multiple birth dates in EAC-CPF? How might you tie those assertions to a source?

Use Case 2

EAC-CPF makes it very easy to attribute multiple names to a single identity. But how would you provide evidence about why a specific name form was included? When the author known as Ntozake Shange published the first edition of her first play, for instance, her first name was written as Ntosake, not Ntozake.

Use Case 3

This next use case is completely fabricated (although I am sure that very real examples exist that could be used in its place), but let us imagine that we have conflicting reports about both the location and the date of Hester Thrale's death. How might we represent and cite that information in EAC-CPF?

Use Case 4

For our last use case, consider that the Connecting the Dots project decided to record that in addition to being a friend of Hester Thale's family, Sir Josuha Reynolds was also related to Hester Thrale since her family employed him to create a series of paintings for the family library. How was that relationship encoded for Connecting the Dots project? How might that relationship be encoded to include information about the archival evidence used to support that claim?

Suggested Approach

Description

For the suggested approach, which should certainly be refined based on community input, I rely heavily on the use of attributes to connect disparate parts of an EAC-CPF record. This approach is essentially implied by the feature request as well as the fact that EAC-CPF already requires that every source that is used in the construction of the EAC record be added to the "control" section of the record, rather than being encoded in situ, repeatedly in the record. Although there are nearly endless options for how we could name new elements and attributes, as well as how to model those nodes, I will not go into exacting detail in this paper. Instead, I expect that those decisions would be worked out after a general approach has been selected.

This particular approach would allow users to connect a new element, temporarily named “evidence”, that would be required (or, at least expected) to be connected to a “source” element and could optionally be connected to a “maintenanceEvent” element (both of which are part of EAC’s control section, and could thereby be included EAD3, if desired). In my testing, this would likely require adding an additional three elements and six attributes. As an overview, the new nodes that I have experimented with while using with this approach are:

- evidence (0 to unbounded)
    - @fromSource (required)
    - @supportsAssertion (optional. can be used to specify the exact data element)
    - @associatedMaintenanceEvent (optional. can be used to associate an agent/date/note)
    - citedRange (0 to unbounded)
        - @unit (optional. can be used to specify where the cited range of materials represents a page number, volume number, etc.)
    - foundData (0 to 1; can only contain text, which is why a new element name is suggested rather than repurposing something like descriptiveNote)
- @assertionRank (optional. added to a parent element of ‘evidence’; can be used when assertions of the same type contradict)
- @relatedAssertion (optional. added to a parent element of ‘evidence’; can be used when assertions of the same type contradict)

So, what might the encodings look like using this approach for each of our four use cases?...

Sample Encoding

Use Case 1

The maintenance event, where the agent and date information would reside:

```
<maintenanceHistory>
  ...
  <maintenanceEvent xml:id="me1">
    <eventType>revised</eventType>
    <eventDateTime standardDateTime="2019-07-14T13:23:19"/>
    <agentType>human</agentType>
    <agent>M. E.</agent>
    <eventDescription>A note describing the fact that a user added a new assertion, if we so choose to include one as part of the update process. Keep in mind that one maintenance event could include multiple assertions, though.</eventDescription>
  </maintenanceEvent>
  ...
```

```
</maintenanceHistory>
```

The source, from which the evidence is cited:

```
<sources>
  <source xml:id="ref1" xlink:href="https://www.theguardian.com/books/2018/aug/10/langston-hughes-born-a-year-before-accepted-date-poet">
    <sourceEntry>Langston Hughes 'born a year before accepted date', researcher finds.</sourceEntry>
  </source>
...
</sources>
```

Encoding the two birth dates, with the second one including the new “evidence” element:

```
    <existDates assertionRank="preferred">
      <dateRange>
        <fromDate localType="http://socialarchive.iath.virginia.edu/control/term#Birth" standardDate="1902-02-01"/>
        <toDate localType="http://socialarchive.iath.virginia.edu/control/term#Death" standardDate="1967-05-22"/>
      </dateRange>
    </existDates>

    <existDates assertionRank="deprecated">
      <dateRange>
        <fromDate xml:id="fdnew" localType="http://socialarchive.iath.virginia.edu/control/term#Birth" standardDate="1901-02-01"/>
        <toDate localType="http://socialarchive.iath.virginia.edu/control/term#Death" standardDate="1967-05-22"/>
      </dateRange>
      <evidence fromSource="#s1" supportsAssertions="#fdnew" associatedMaintenanceEvent="#me1" localType="http://purl.org/spar/cito/isSpeculatedOnBy">
        <foundData>Multiple newspaper articles in 1901 reference Langston Hughes and his mother prior to Langston's established birth year of 1902.</foundData>
      </evidence>
    </existDates>
```

As illustrated above, for this particular use case to work, it would be ideal if EAC-CPF allowed the existDate elements to be repeatable (the only way to make this work without changing that aspect of the EAC schema would be to add a dateSet to the lone existDates element, which does not seem accurate). Further, given that existDates *can* repeat in ISAAR(CPF), this seems like a good addition to EAC-CPF regardless of this specific feature request.

Use Case 2

```
...
  <maintenanceEvent xml:id="me2">
    <eventType>revised</eventType>
    <eventDateTime standardDateTime="2019-07-15"/>
    <agentType>human</agentType>
    <agent>M. E.</agent>
  </maintenanceEvent>
 </maintenanceHistory>
 <sources>
  <source xml:id="s2" xlink:href="http://hdl.handle.net/10079/bibid/1164279">
   <sourceEntry>For colored girls who have considered suicide, when the rainbow is enuf / Ntosake
[sic] Shange ; [drawings by Wopo Holup].</sourceEntry>
  </source>
 </sources>
</control>
<cpfDescription>
 <identity>
  <entityType>person</entityType>
  <nameEntry>
   <part>Shange, Ntozake.</part>
  </nameEntry>
  <nameEntry>
   <part>Shange, Ntosake</part>
   <evidence fromSource="#s2" associatedMaintenanceEvent="#me2">
    <citedRange unit="page">title page</citedRange>
   </evidence>
  </nameEntry>
  <nameEntry>
   <part>Williams, Paulette L. 1948-</part>
  </nameEntry>
 </identity>
 ...
```

Note the use of the citedRange element in this case to indicate where exactly the evidence was found on the source.

Use Case 3

Maintenance event information:

```
    <maintenanceEvent xml:id="me3">
```

```
        <eventType>revised</eventType>
        <eventDateTime standardDateTime="2019-07-17"/>
        <agentType>human</agentType>
        <agent>M. E.</agent>
        <eventDescription>This is a note about why new data was added. It could include information about why an assertion is now marked as deprecated, etc.</eventDescription>
      </maintenanceEvent>
```

Source information:

```
      <source  xml:id="s3" xlink:href="http://www.worldcat.org/oclc/47811007">
        <sourceEntry>Fanny Burney: A Biography</sourceEntry>
        <!-- note how descriptiveNote here is being used as a citation.   Also, why does EAC have a citation element if it does not have any structure? -->
        <descriptiveNote>
          <p>Harman, Claire. 2001. <span localType="title">Fanny Burney: a
            biography</span>. New York: Alfred A. Knopf. </p>
        </descriptiveNote>
       <!-- See objectXMLWrap as a better example, using TEI to encode the citation... but other things like CSL, etc., could also be used. -->
        <objectXMLWrap>
          <biblStruct xmlns="http://www.tei-c.org/ns/1.0">
            <monogr>
              <author>
                <persName>
                  <forename>Harman</forename>
                  <surname>Claire</surname>
                </persName>
                <idno type="scopus">37083592200</idno>
                <idno type="lcaf">http://id.loc.gov/authorities/names/n83006657</idno>
              </author>
              <title level="m">Fanny Burney: A Biography</title>
              <imprint>
                <pubPlace>New York</pubPlace>
                <publisher>Alfred A. Knopf</publisher>
                <date when="2001-08-21"/>
              </imprint>
            </monogr>
          </biblStruct>
        </objectXMLWrap>
      </source>
```

An approach to encode conflicting statements:

```
<place assertionRank="deprecated" xml:id="p1" relatedAssertion="#p2">
  <placeRole>death</placeRole>
  <placeEntry vocabularySource="lcsh">Bristol (England)</placeEntry>
  <date standardDate="1821">1821</date>
</place>

<place assertionRank="preferred" xml:id="p2" relatedAssertion="#p1">
  <placeRole>death</placeRole>
  <placeEntry vocabularySource="lcsh" xml:id="pe1">Newport (England)</placeEntry>
  <date standardDate="1822" xml:id="dd1">1822</date>
  <evidence fromSource="#s3" supportsAssertion="#dd1"
associatedMaintenanceEvent="#me3">
    <citedRange unit="page">312</citedRange>
  </evidence>
  <evidence fromSource="#s3" supportsAssertion="#pe1"
associatedMaintenanceEvent="#me3">
    <citedRange unit="page">313</citedRange>
  </evidence>
</place>
```

Note here the use of the relatedAssertion attribute to unambiguously link the sibling place elements. This would certainly not be required, but it could be helpful for constructing user interfaces, or visualizing the data, especially when the assertionRank attributes were absent or set to a value like “normal” instead of “preferred”.

Use Case 4

Source information (with no maintenance event, since it would be optional to include a reference to an agent and date):

```
<source xml:id="s4" xlink:href="http://nrs.harvard.edu/urn-3:FHCL.Hough:hou00264">
  <sourceEntry>Hester Lynch Piozzi Manuscripts, 1765-1820.</sourceEntry>
</source>
```

Evidence for the relationship:

```
<!-- friendOf relationship, with no evidence -->
<cpfRelation cpfRelationType="associative" xlink:actuate="onRequest"
  xlink:arcrole="http://nrs.harvard.edu/urn-3:FHCL.Hough:12242299"
  xlink:href="http://hdl.handle.net/10079/k0p2nt3">
  <relationEntry>Reynolds, Joshua, Sir, 1723-1792</relationEntry>
  <descriptiveNote>
```

```
            <p>Friend and frequent guest of the Thrales.</p>
          </descriptiveNote>
        </cpfRelation>
        <!-- employedBy relationship, with evidence -->
        <cpfRelation cpfRelationType="associative" xlink:actuate="onRequest"
          xlink:arcrole="http://nrs.harvard.edu/urn-3:FHCL.Hough:12242301"
          xlink:href="http://hdl.handle.net/10079/k0p2nt3">
          <relationEntry>Reynolds, Joshua, Sir, 1723-1792</relationEntry>
          <evidence fromSource="#s4">
            <foundData>The letters found in folders 1-5 detail that the Thrales ordered a series of 13
paintings from him for their library.</foundData>
            <citedRange unit="box">1</citedRange>
            <citedRange unit="folder">1-5</citedRange>
          </evidence>
        </cpfRelation>
```

Alternative Approaches

There are a variety of alternative approaches that could be taken, although each would either result in encoding even more duplicative data (e.g. following an approach similar to using an “evidence” element, but also allowing a source/citation element as well as an element such as “assertionEvent” to be encoded as a child element in order to capture information about the agent who made the assertion) or would result in achieving fewer of the original recommendations (e.g. using descriptiveNote -- which often is already used to include evidence for why an assertion has been made -- or a similarly-modelled element without worrying about connecting that assertion to an agent). These approaches, however, would not be helpful for encoding conflicting assertions without similar extensions already described with the suggested approach. Whether or not that is deemed important is still an open question.

Additional Background Information

In the process of considering this proposal, I first reviewed what was already possible in both EAC-CPF and EAD in terms of encoding citations. I did not proceed too far with that investigation however, since the Shared Schema subteam (of which I am part) is in the process of writing a report that is comparing the similarities and differences between the data models currently expressed in EAC-CPF and EAD3.

Given that the results of the Shared Schema subteam will likely inform the direction of EAC and EAD, I did not focus on the current state of those standards, nor which direction they might head in the future (i.e. whether the schemas will share data models, or whether they will remain separate). I should note, however, that EAD3 includes an element (not present in previous

versions of EAD) named “footnote”, which has no equivalent in EAC-CPF. The EAD3 tag library instructs that this element should be used “to annotate text to indicate the basis for an assertion or citing the source of a quotation or other information.”[1] The footnote element can be utilized in EAD3 to accommodate R3 (see above), as well as a few of the other requested updates, it cannot do so very easily or elegantly. Further, the selection of the term “footnote” in EAD3 -- which puts the focus on a specific type of potentially-unparsable note element that is used within printed and electronic documents that are meant to be read in a linear fashion -- did not seem to make that element name a very good fit for EAC-CPF, outside of some of the elements provided in EAC’s description section.

Outside of EAS, I also looked at sampling of other standards to see how they represent similar types of evidence-based assertions, including TEI (https://tei-c.org/), MARC Authorities (specifically https://www.loc.gov/marc/authority/ad670.html), CiTO (https://sparontologies.github.io/cito/current/cito.html), the Scientific Evidence and Provenance Information Ontology (https://github.com/monarch-initiative/SEPIO-ontology), and Wikidata’s data model for what it refers to as “snaks” (https://www.mediawiki.org/wiki/Wikibase/DataModel#Snaks). SEPIO, in particular, provides a useful diagram to consider, as seen here from a common root diagram for what they define as an “assertion”, which lines up reasonably well with Daniel’s request, although the modelling in SEPIO is certainly more complicated: https://github.com/monarch-initiative/SEPIO-ontology/wiki/Assertion

I should also add that I was specifically interested in exploring this feature request since I was once told that the digital project Syriaca.org (http://syriaca.org/) was interested in using EAC-CPF, but decided to use TEI exclusively instead since EAC could not connect assertions with sources of evidence. Here is an example of what the Syracia project is encoding in TEI that they cannot encode in EAC-CPF:

> TEI relation element from Syraica.org, from http://syriaca.org/person/3/tei:
>
> ```
> <relation name="born-at" active="http://syriaca.org/person/3"
> passive="http://syriaca.org/place/78" source="#bib3-2">
>    <desc xml:lang="en">This author was born at a known location.</desc>
> </relation>
> ```
>
> This relationship points to the following source in the TEI header by means of an xml:ID/xml:IDREF relationship:
>
> ```
> <bibl xml:id="bib3-2">
>    <title level="m" xml:lang="en">The Scattered Pearls: A History of Syriac Literature and
>    Sciences</title>
>    <ptr target="http://syriaca.org/bibl/4"/>
> ```

[1] Encoded Archival Description Tag Library Version EAD3: https://www.loc.gov/ead/EAD3taglib/index.html#elem-footnote

```
    <citedRange unit="entry">3</citedRange>
    <citedRange unit="pp">225</citedRange>
</bibl>
```

This example accommodates R3, R4, and R6 from Issue #41. This example also fulfills another scholarly need by making use of the “citedRange” element in TEI. Although that would be possible to do in EAC-CPF by using a descriptiveNote element within a source element (assuming the revisions included an update to allow sources to be linked to assertions), that does not provide an unambiguous place to encode information about where within a source the citation refers, which is why the TEI approach is more attractive. That said, EAC-CPF’s data model for “source: could be expanded to include a citedRange element. But even if it were, it would not change the fact that EAC still has different elements for “source” and “citation”, whereas TEI instead has different elements that can be used to express different levels of citations (e.g. bibl, biblStruct, etc., as described at https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-model.biblLike.html).

In summary, what is most clear to me is that there is no way to connect “assertions” (i.e. statements encoded in EAC-CPF) to “sources”, even though there is a clear mechanism in EAC’s control section to provide a list of sources. I hope that the above sections, as well as the simple TEI example, indicate that this, by itself, should be a relatively simple problem to address should we choose to include it in the EAC revision. However, I think that there are additional considerations within the larger EAC and EAD schemas that should be tackled at the same time, which certainly could influence any approach taken during the process of adding evidence-based assertions to the EAS family.

Additional Considerations

Outside of the request to add assertion-level metadata to EAC, there are a number of elements in EAD and EAC that could be reviewed in light of their data models and how they do (or do not) meet current needs for citing sources. EAD3, for instance, has bibref, archref, and ref, which differ only in the name of their element and where they are permitted (e.g. you cannot use bibref in a paragraph, and you cannot use a ref or archref element in a bibliography, but each serve the same purpose). Similar models exist for note, controlnote, footnote (newly added to EAD3), didnote, etc. Additionally, the source element in EAD/C is more nuanced than the citation element in EAC, which is not available in EAD (outside of the control section, that is). Although outside the scope of this overview, those additional data models (and schema inconsistencies) should continue to be considered during the major revision of EAC.

An incomplete list of thoughts so far include:

- Consider changing EAD @source to @vocabularySource. It would be nice to use @source the same way that TEI uses @source, but current practice in EAD does not allow this.

- Consider selecting an attribute or method other than @localType to encode references to external ontologies. Outside of xlink:arcrole (which is not widely available), it seems that @localType is being used to fill this void.
- Remove “footnote” as an EAD3 element, or model it differently so that it could be shared amongst EAS schemas.
- How to handle “citation” vs. “source”, especially since “citation” cannot be a child of source, and yet that’s often how source/descriptiveNote is used in existing EAC records? Also see the definitions in the EAD/C tag libraries.
- Whatever data model is selected for evidence-based assertions, that choice will be impacted by the decision regarding plural/singular sections in EAC’s Description section. Given that, the plural/singular choice should probably be decided upon first.
- Other outliers that need some sort of evidence or citation option, which don’t already have descriptiveNote or citation:
    - nameEntry
    - useDates
    - existDates (also needs to be unbounded)
- Places that we would likely exclude for evidence-based assertions, even though those data models include desriptiveNote and/or citation:
    - control
    - alternativeSet